This paper was submitted to the 10th International Conference of the International Society for Scientometrics and Informetrics, Stockholm, Sweden, July 24-28, 2005

Visualizing 60 Years of Anthrax Research

Steven A. Morris* and Kevin W. Boyack**

*samorri@okstate.edu
Electrical and Computer Engineering, Oklahoma State University
202 Eng. So., Stillwater, Okla. 74078, USA

**kboyack@sandia.gov
Sandia National Laboratories, P. O. Box 5800, MS-0310, Albuquerque, NM 87185, USA

Abstract: Using a collection of 2472 papers covering 60 years of anthrax research, we demonstrate three techniques for visualizing and mapping knowledge in a scientific specialty. Timelines, maps of clusters of papers by time, reveal the temporal changes in the specialty. Crossmaps, maps of the correspondence of groups of entities of different entity-types, reveal overlapping relations among sub-specialties within the specialty. Usage plots, maps of entity usage over time, allow visualization of emergence and obsolescence in the specialty of key entities such as seminal references and important researchers. The timeline visualization of the anthrax collection of papers reveals that anthrax research grew in two distinct phases, and that the specialty experienced major disruptions in reaction to the Soviet anthrax bioweapons accident at Sverdlovsk in 1979, and the anthrax postal bioterror attacks in the United States in 2001. A crossmap of research fronts to reference clusters reveals groups of key references in the specialty, and the overlapping relation of those groups to research topics in the specialty. A usage plot of references reveals the temporal emergence and obsolescence of key groups of references in the specialty.

Introduction

A collection of papers that covers a specialty constitutes a collection of research reports, vetted by the review process. Study of such a collection permits subject matter experts to assess the state of the art of the specialty and present that assessment to industry leaders and policy makers for decision making purposes. The motivation for the work presented here is the need to efficiently present important information about a research specialty in various ways to subject matter experts that are monitoring the specialty. The three techniques described here, timelines, crossmaps, and usage plots, are particularly applicable to bibliometric analysis (White & McCain, 1989) and knowledge mapping (Borner, Chen, & Boyack, 2003), and provide a general method of visualization of complex structure of a specialty.

The example presented here is on the topic of anthrax research and covers a period of about 60 years. An initial study on anthrax research was used by Morris et al. (2003) to show the use of bibliographic coupling to form research fronts of papers. Research fronts were defined as groups of papers that tend to cite common references. Such groups of papers tend to cover a common research sub-topic in the specialty. Morris et al. showed that timelines of research fronts can be used to visualize structure and dynamic changes in a research specialty. The collection of papers on anthrax studied in that previous work was updated and the analysis of the updated collection is presented here.

Background on anthrax research

Early anthrax research covers a time period from 1946 to about 1975 and covers toxin research, vaccines, inhalational anthrax and medical treatment. After a hiatus of research in the 1970's, several research fronts emerged with varied growth characteristics. Vaccine and gene sequencing research fronts have proceeded steadily for 15 to 20 years, for example, while research on anthrax toxins shows a pattern of rapid growth and specialization. The topic of anthrax bioterrorism emerged since 1999 in response to perceived threats, and the Fall, 2001 bioterror attacks through the U. S. postal services have produced a shock to the specialty

that generated great interest in anthrax research and produced new research fronts dealing with aspects of anthrax bioterrorism.

Acquisition and storage of data

Morris et al. (2003) describe the method used on December 12, 2001 to acquire the previous dataset of 833 papers on anthrax research from ISI's Web of Science product. The collection was updated on February 25, 2003 using the search term "anthrax OR anthracis" to capture papers available from the Web of Science from 1945 forward. When this collection was combined with the previous collection and duplicates removed, the anthrax collection totaled 2472 papers. Figure 1 diagrammatically lists the number of entities and links in the collection.

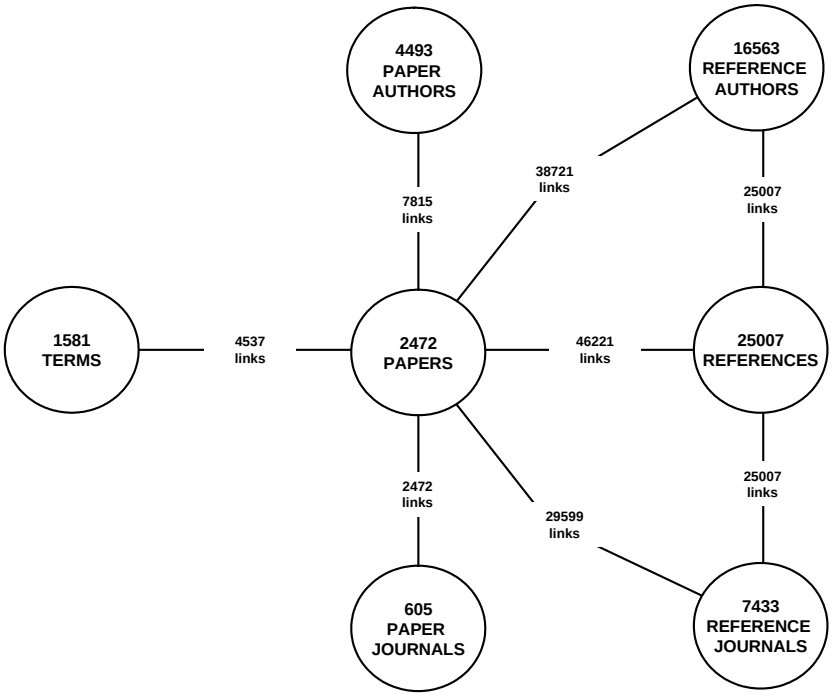

Figure 1. diagram showing the number of entities and links in the anthrax collection.

The balance of this paper starts with a demonstration of three techniques, timelines, crossmaps, and usage plots, to characterize both the history and current state of the field of anthrax research using the updated anthrax paper collection. This is followed by discussions of the effects of postal bioterror attacks and research funding on anthrax research, and by a summarization of the benefits of using visualization techniques such as those detailed here.

Timeline: Research fronts

The paper to reference matrix consists of 2472 papers having 46221 links to 25007 references. Papers that did not have at least five common references with another paper were discarded, leaving 987 papers, which were clustered into 35 research fronts. Similarity values between papers were calculated using bibliographic coupling counts (Kessler, 1963) that were converted to similarities using the cosine coefficient (Salton, 1989). Distances between pairs of papers were found by subtracting their similarity from unity. Clustering was performed using hierarchical agglomerative clustering with incremental sum-of-squares (Ward's method) linkage (Gordon, 1999). A dendrogram seriation routine was used to place related research fronts close to each other on the tree (Morris, Asnake & Yen, 2003). Examination of the 35 research fronts showed that the papers generally fell into coherent research topics and did not show any off-topic clusters of papers that needed to be discarded.

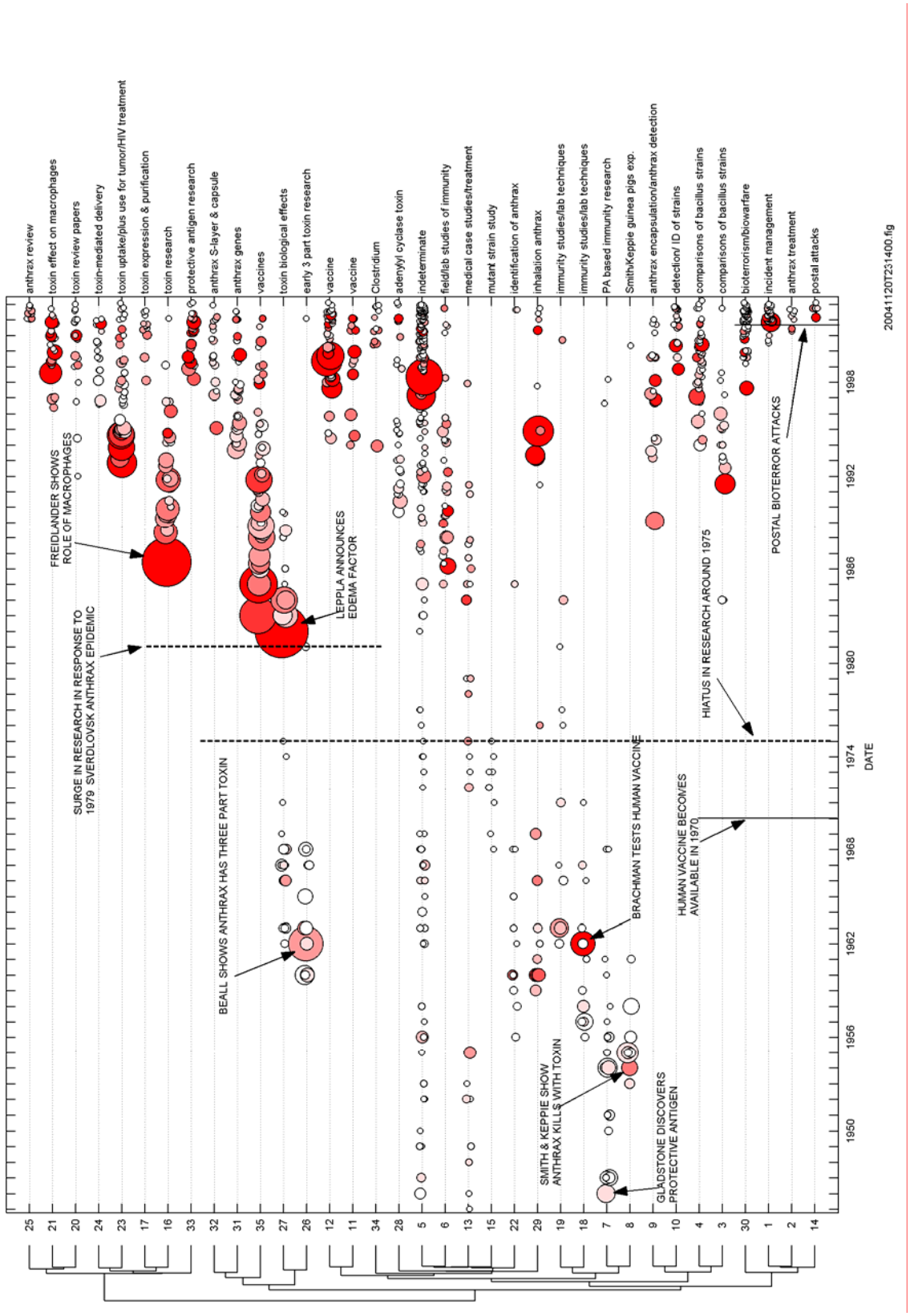

Figure 2. Research front timeline for the anthrax case study.

Figure 2 shows a research front timeline for the anthrax collection. The figure is rotated 90 degrees to allow a higher resolution view. Looking at the left, the identifying numbers for each research front are printed in a column to the right of the clustering dendrogram. The

papers in each research front are plotted by time in horizontal tracks, with the research front labels on the right side of the plot. Research front labels were manually assigned after browsing the titles of the papers in each research front for themes. The circles on the plot correspond to papers and the size of each circle is proportional to the number of times that the paper was cited. Each circle is shaded in proportion to the number of times its corresponding paper has been cited in the last year of the collection (February 2002 to February 2003). The timeline has been marked up manually to assist the viewer in understanding the interpretation to follow.

It is easy to see from the timeline that the collection falls into two distinct sections 1) research from 1945 to about 1975, and 2) research from 1976 to 2003. In general, early research fell into three categories: toxin research, medical treatment, and vaccine research, while the more current research shows signs of increased specialization. It seems probable that the great surge in the number of research papers starting in the 1990's is tied to government funding of research in response to bioterrorism threats. The research fronts can be classified as follows:

Early research (1945 to 1975)

- Research fronts 7 and 8 are the earliest research and cover early immunity studies and Smith and Keppie's seminal guinea pig experiments that showed that anthrax kills with a toxin.
- Research fronts 19 and 18 cover vaccine.
- Research front 29 deals with medical treatment of inhalational anthrax and may be tied to funding of bioweapons research in the 1960's. This work is dormant throughout the 1970's and 1980's, but picks up again in the 1990's, an event that may be tied to government funding of research on bioterrorism in the late 1980's. The key papers in this research front are currently being heavily cited in response to the postal bioterror attacks and resulting intense interest in treating inhalational anthrax.
- Research fronts 15 and 22 are papers that discuss identification of anthrax and discrimination of anthrax strains.
- Research front 13 consists of papers on medical case studies. This research front continues up to the 1990's, with few papers during the late 1950's and 1960's. It is finally superceded around 1985 by research front 6.
- Research fronts 27 and 26 are papers that continue Smith and Keppie's experiments on the anthrax toxin. These papers establish the knowledge on the three-part anthrax toxin until they are superceded by the seminal works of Leppla in 1982 and Freidlander in 1988.

Current research (1976 to 2003)

- In 1979, an anthrax bioweapons accident at Sverdlovsk, in the Soviet Union, may have been the motivating factor for research funding that produced a resurgence of anthrax research in the 1980's (Turnbull, 1991).
- Research fronts 25 to 33 (the top 8 fronts in Figure 2) generally cover the topic of anthrax toxin research. These include research on the three parts of the toxin: protective antigen, edema factor, and lethal factor.
- Research fronts 24 and 23 cover research on using protective antigen to inject materials into cells to induce immunity to AIDS and other diseases.
- Research fronts 32, 31, and 35 cover general anthrax research in the 1980's, anthrax gene sequencing, and the anthrax capsule.
- Research fronts 12 and 11 cover current research in anthrax vaccines.

- Research fronts 34 and 28 are miscellaneous topics. Clostridium is a bacteria that produces botulism toxin and is related to anthrax through bioterrorism research; adenylyl cyclase toxin is closely related to edema factor research.
- Research fronts 9, 10, 4, and 3 cover research on various methods of sensing anthrax and classifying anthrax by strains.
- Research fronts 30, 1, 2, 14, and 25 are bioterrorism related. Research front 30 covered general bioterror research until the postal terror attacks. The postal attacks induced four other research fronts: 1) Research front 1 on incidence management, 2) Research front 2 on treatment of the disease, 3) Research front 14 covering the postal attacks themselves, and finally 4) a series of reviews on anthrax toxin research which was clustered at the top of the figure into the research fronts covering toxin research.

Crossmap: Analysis of references

A review of crossmap visualizations can be found in Morris and Yen (2004). In general, a crossmap shows the correlation between entities of two different types (for example. research fronts and references) in a way that highlights the strengths of individual correlations.

For the anthrax collection, we desired to show the correlation between key references and research fronts. To do this, references that were cited 40 times or more were retained. This yielded 70 highly cited references. Similarities between references were calculated from co-citation counts (Small, 1973) converted to similarities using the cosine coefficient. Distances were computed by subtracting similarities from unity, and clustering was done using hierarchical clustering with Ward's method linkage. A dendrogram seriation routine was used to place references with high similarity close together on the dendrogram.

Figure 3 shows a crossmap of research fronts to references. References are mapped as columns in the crossmap, with a dendrogram at the top of the figure and reference labels at the bottom of the figure. Research fronts are mapped as rows on the crossmap, with clustering dendrogram on the left and research front labels on the right. These research fronts and the dendrogram on the vertical axis match those of the timeline of Figure 2. Given research front *i*, and reference *j*, the size of the circle on the map at row *i*, and column *j*, is proportional to the percentage of papers in research front *i* that cite reference *j*.

In this map rectangular boxes have been manually added to the crossmap to visually highlight significant relations of groups of research fronts to clusters of references. It is easy to see the overlapping correspondence of reference clusters to research fronts:

- At the bottom left of the map, note a series of references that are key references for anthrax bioterrorism. This group of references starts with reference number 71 and ends with reference number 34 as seen at the top of the map. Reference 71, Jernigan 2001, is a notable reference that reports on 10 cases of inhalation anthrax from the postal bioterror attacks. Inglesby, 1999, reference 3, is a policy paper on anthrax bioterrorism.
- At the far left is a series of four references: Sambrook 1989, Ash 1991, Keim 2000, and Keim 1997, which are key references for methods of detecting anthrax and discriminating among strains.

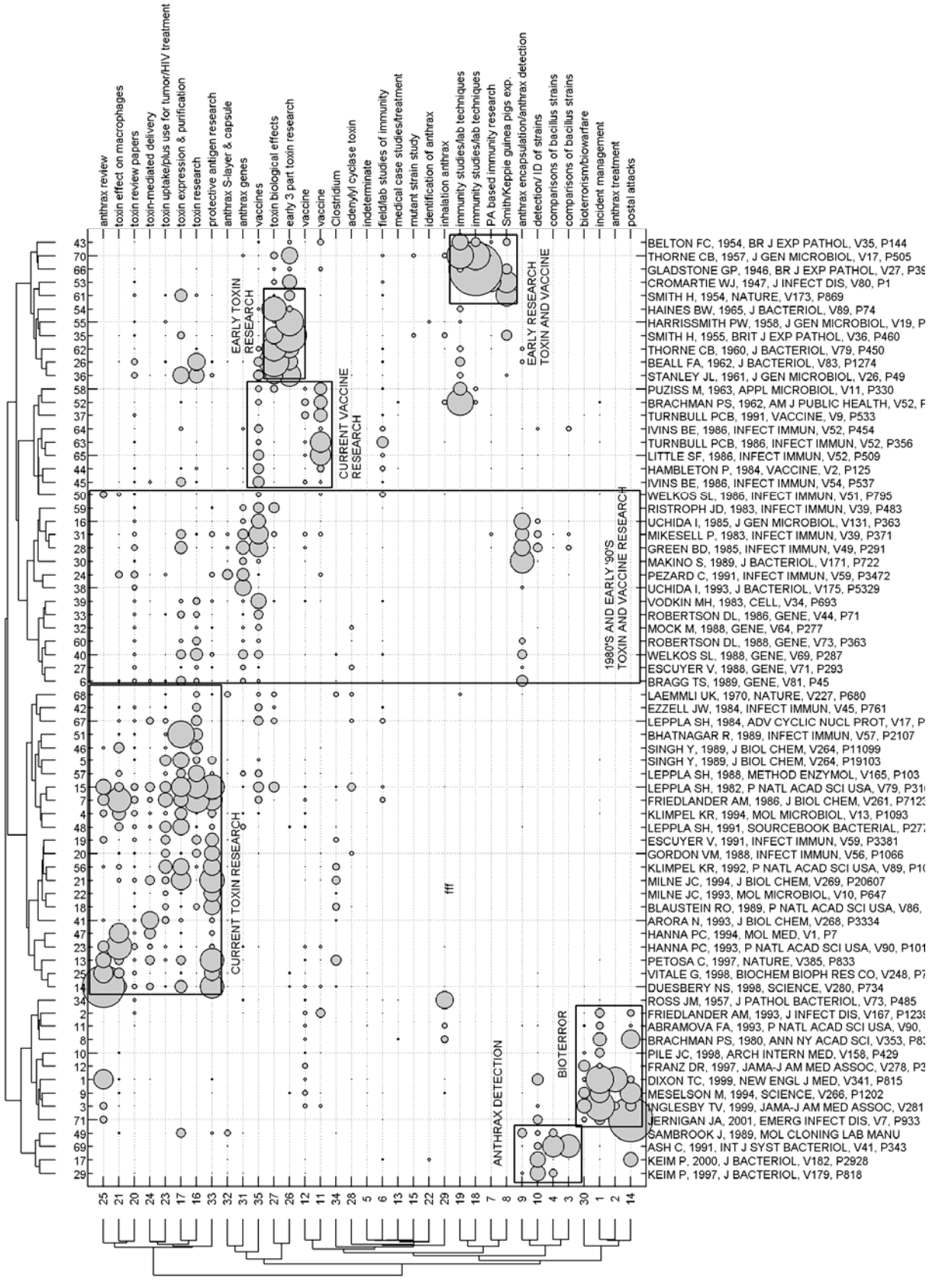

Figure 3. Research front to reference crossmap for the anthrax collection.

- At the top of the map a series of 23 references comprise the key references on the topic of toxin research. This group starts with reference 14 on the left and ends with reference 68 on the right. There is a great deal of overlap in the correspondence of these references to different research fronts in toxin research (research front 25 at top down to research front 33). The key references for all of toxin research are Leppla 1982 and Friedlander 1986, references 15 and 7 respectively. Another important reference is Duesbery 1998, reference 14, corresponding to a paper that explained the mechanism by which anthrax lethal factor kills cells.
- In the center of the map, references from 6 on the left to 50 on the right, correspond to papers published in the 1980's and 1990's in anthrax before a lot of specialization occurred. These are cited from many research fronts, but particularly from research fronts on vaccines and anthrax genetic sequencing.
- At the center right of the map, a series of references (from 45 on the left to 58 on the right) are key references for current vaccine research. Among these is Brachman 1962, reference 52, which corresponds to a report on efficacy of the anthrax vaccine that is commonly used today.
- On the right of the plot are references that were used by papers in research fronts for early anthrax research from 1945 to 1975. The group of 5 references at the extreme right corresponds to the earliest research, and includes Smith and Keppie's original study that showed that anthrax kills with a toxin, Smith 1954, reference 61. Immediately to the left of this group are references, from reference 36 on the left to reference 54 on the right, used by papers in the research front from the 1950's that established that the anthrax toxin has three parts.

Usage plot: Reference usage

A usage plot is a specialized type of crossmap which uses time on the y-axis in place of an entity type. Figure 4 shows a map of reference usage for the anthrax collection. In this plot the references arrayed on the x axis are identical to those from the research front to reference crossmap, Figure 3. The rows correspond to paper years. Given year *i* and reference *j*, the size of a circle at row *i* and column *j* on the map is proportional to the number of times that reference *j* was cited in year *i* by all papers in the dataset, irrespective of research front. The main purpose of this map is to show obsolescence of references. Because of the small volume of papers in the early research period from 1946 to 1975, the size of the circles in this period are magnified 4 times over the sizes in the later period. The following features are visible on this map:

- On the extreme right the series of references from 62 on the left to 43 on the right have become obsolete. They cease to be cited around 1968, a year which may have corresponded to cuts in funding of anthrax research. These references are not cited much even after anthrax research picks up again in the late 1980's and early 1990's.
- Two references from early research, Stanley 1962 and Beall 1962, references 36 and 26, correspond to papers that characterize anthrax lethal factor toxin. As shown on the plot, these references are still current and being cited.
- Reference 34, Ross 1957, corresponds to a paper on how inhalational anthrax develops in the lungs. After a long period of no citation that started about 1966, this reference is being cited heavily since the postal bioterror attacks because of the current intense interest on treating inhalational anthrax that resulted from those attacks.

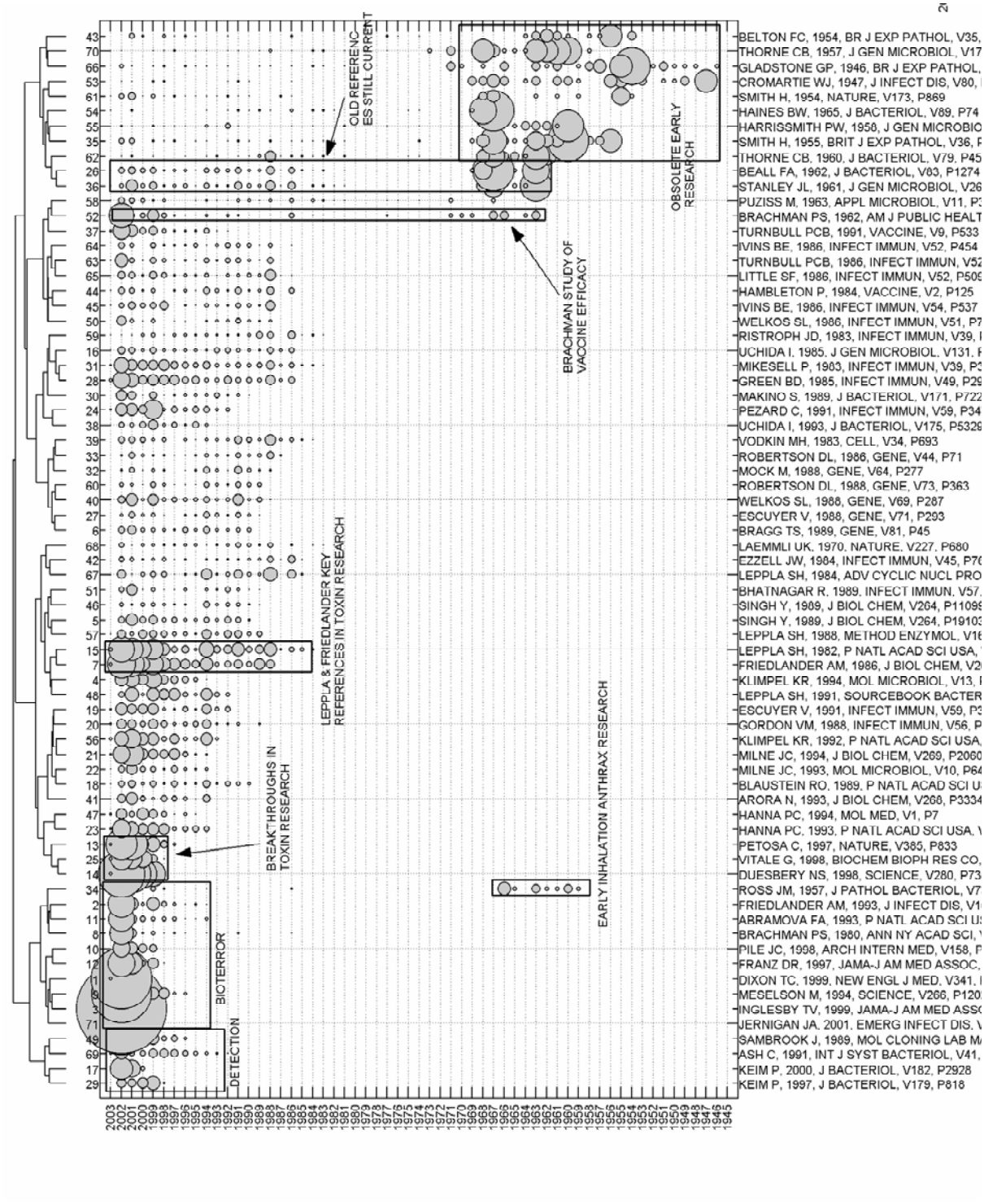

Figure 4. Reference usage plot for the anthrax case study.

- Note that it is easy to see the emergence and use of key references on the detection of anthrax and discrimination of strains of anthrax. These are the references on the extreme left (starting with reference 29 on the left to reference 49 on the right), and the initial reference in this group is Ash 1991 (reference 69).
- Note the two seminal references for toxin research are in the center of the plot. These are Leppla 1982 and Friedlander 1986, reference 7 and 15 respectively. Leppla corresponds to a key paper on edema factor, while Friedlander announces the discovery of the role of macrophages in the spread of the disease within a host. While

the Leppla paper appears in 1982, it does not start to be heavily cited until 1988, a 6 year delay. The year 1988 may be a year in which government funding of anthrax research was increased dramatically in response to bioterror threats.

Finally, note the great number of citations received by key bioterror references in the year 2002. These are to the left of the plot, from reference 71 on the left to reference 34 on the right. This heavy number of citations reflects the intense interest in anthrax bioterror after the postal attacks in late 2001.

Discussion of postal bioterror attacks

The postal bioterror attacks in Fall, 2001, caused a great shock in the specialty of anthrax research. The previous study by Morris et al. (2003) that was conducted on papers gathered on December 23, 2001, about two months after the attack, showed that 6 papers had already been published in reaction to those attacks. Figure 5 is a timeline from that study that shows a research front (number 10) on anthrax bioterrorism. In this diagram the citation links between papers are shown on the timeline to papers cited heavily from the bioterrorism research front. Six papers that appeared after the bioterror attacks are shown and it is noted that five of these six papers cited a paper by Dixon that covered the treatment of anthrax. Previous to this, papers in the bioterror research front rarely cited Dixon. This indicated that a research front on medical treatment of anthrax was about to emerge.

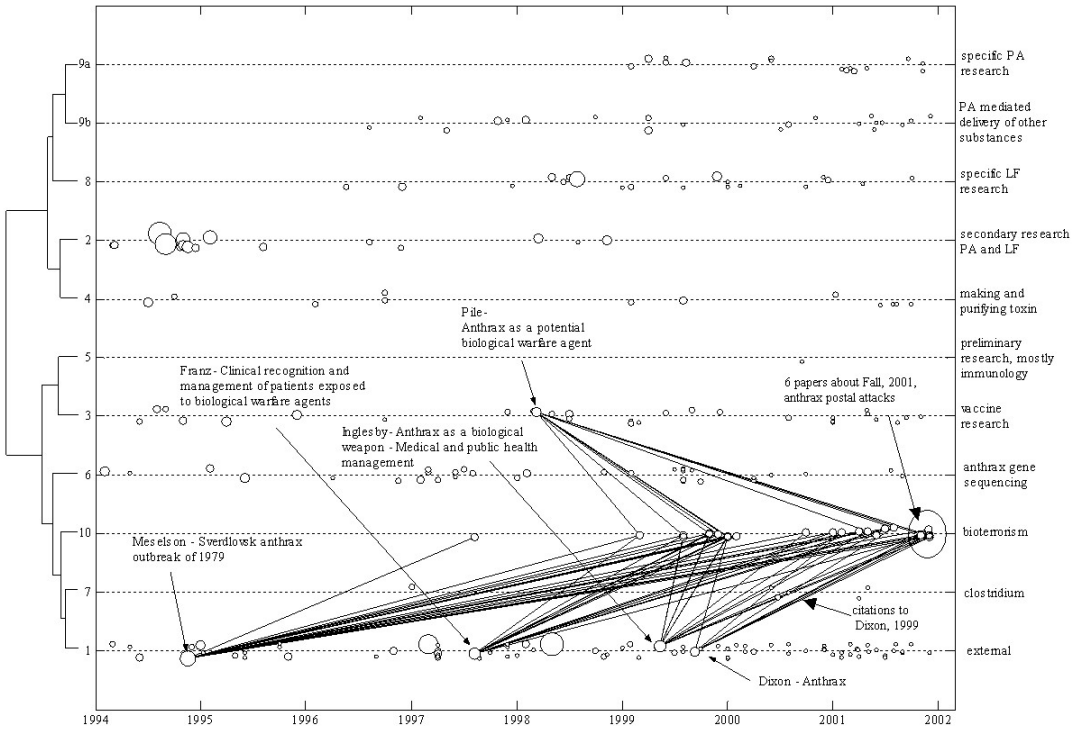

Figure 5. A timeline from an earlier study showing the early effects of the postal bioterror attacks on the literature.

Looking at Figure 2, which is based on papers gathered on February 25, 2003, about 14 months after the bioterror attacks, the response of the specialty to the postal attacks is evident.

There are three additional bioterror related research fronts in the specialty: 1) research front 1 dealing with bioterror incidence management, 2) research front 2 dealing with medical treatment of anthrax, and 3) research front 14 dealing specifically with the postal attacks themselves. Because of the great interest in anthrax research that was generated after the attacks, a series of anthrax toxin research review papers was generated, which became research front 25 at the top of the timeline.

As seen in the timeline of Figure 5, the prediction of a new research front was accurate. There was, however, no anticipation that research fronts on incidence management, the postal attacks themselves, and anthrax toxin review would emerge. Looking at Figure 3 it can be seen that the paper by Dixon on medical treatment of anthrax is heavily cited from the new research front on medical treatment.

Funding of anthrax research

An effort was made to correlate surges of papers as shown in the timeline of Figure 2 to funding records from U. S. government sources. The total number of funded projects per year for two of these sources, the National Institutes of Health (NIH), and the Department of Defense (DoD) are shown in Figure 6. Other funding agencies show similar trends, but with far fewer funded projects. There may be gaps in the funding databases in the years prior to 1990, so it is not possible to correlate funding to the research inactivity that started about 1970 and ended about 1980. Note, however, a surge of funding in the early 1990's that may be tied to research on vaccines, defense against anthrax bioweapons, and Gulf War Syndrome, for which the anthrax vaccine was a suspected cause. Note also the large surge in projects in 2002 and 2003 that was undoubtedly an immediate response to the postal bioterror attacks.

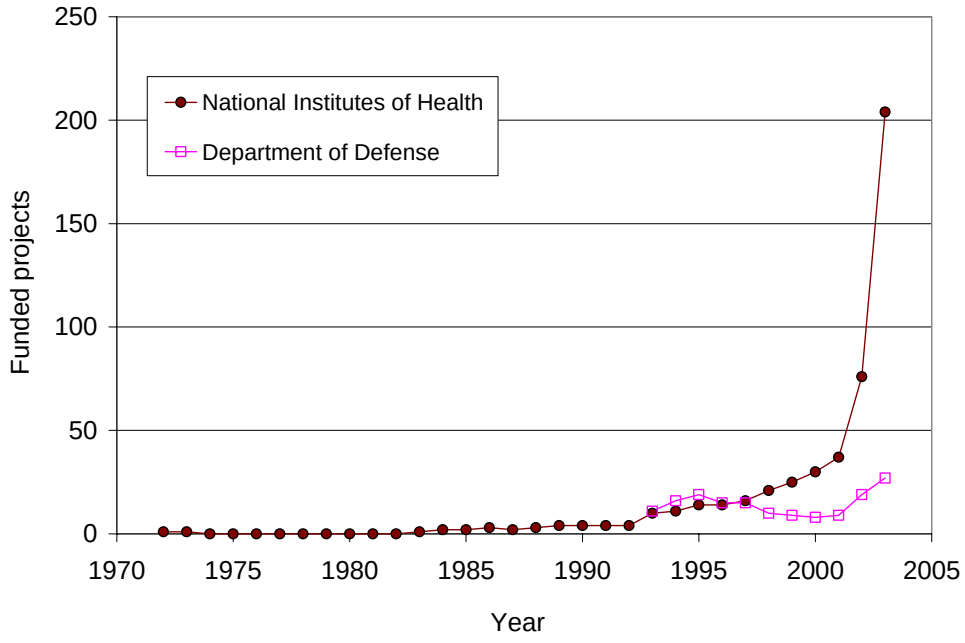

Figure 6. Number of funded anthrax research projects over time.

Final discussion

Given a collection of papers on the topic of anthrax research, the methods and visualizations detailed above were used to generate the following knowledge:

- **Research fronts**. The full collection of documents was separated into 35 separate research fronts, each of which comprises a subspecialty of research within the anthrax research specialty. The time relation of these research fronts was exposed through the timeline of Figure 2 and from this the active and inactive research fronts were identified and catalogued.
- **Reference groups**. The key references within the specialty were identified, clustered into related groups, and their overlapping relations to the specialty's research fronts were shown. The use of these references over time was visualized and the obsolete and current key reference groups were identified. This analysis is important because it allows subject matter experts to quickly identify the seminal references associated with important subtopics in a specialty. Identification of these references allows the subject matter experts to monitor for new papers that cite these key references, and further allows them to educate themselves on the key elements of these subspecialties by reading the papers corresponding to such references. Furthermore, the overlap of these reference groups with multiple research fronts allows subject matter experts to map how research fronts are related.

It is also possible to investigate relations among other entity-types in the anthrax collection using the visualization techniques presented here. Due to space limitations, we do not show visual examples of these relations, but simply mention what can be learned from them.

- **Reference author groups.** Key groups of reference authors in the specialty can be identified, clustered into groups, and their overlapping relations to research fronts in the specialty can be shown in a crossmap. Additionally, a usage plot showing the temporal use of reference authors by research fronts can be visualized, showing how groups of reference authors, corresponding to "schools of thought" in the specialty, emerge, were used, and become obsolete. Furthermore, this visualization and analysis shows which reference authors are currently being used. The visualization shows experts in each research front as those authors that are well cited by the research fronts.
- **Paper author groups**. Visualization of paper author usage and clustering allows the identification of teams of researchers, helps to identify prolific authors and also shows which researchers are currently active in the specialty
- **Term groups**. Visualization of terms and groups of terms is very useful for labeling research fronts.

This case study illustrates that the visualization techniques introduced in this paper can be used to extract a large amount of useful information about a research specialty from a collection of papers that cover that specialty. Research subtopics, seminal papers, important experts, active research teams, and the relations among them are information that is produced that can be made available to subject matter experts to assess the state of the specialty and make recommendations to planners and research managers.

References

Borner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37, 179-255.

Gordon, A. D. (1999). *Classification* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*, 10-25.

Morris, S. A., Asnake, B., & Yen, G. (2003). Optimal dendrogram seriation using simulated annealing. *Information Visualization, 2*(2), 95-104.
Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Timeline visualization of research fronts. *Journal of the American Society for Information Science and Technology, 54*(5), 413-422.
Morris, S. A., & Yen, G. (2004). Crossmaps: visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Science of the United States, 101*(suppl. 1), 5291-5296.
Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Reading: Addison-Wesley.
Small, H. (1973). Cocitation in scientific literature - new measure of relationship between 2 documents. *Journal of the American Society for Information Science, 24*(4), 265-269.
Turnbull, P. C. B. (1991). Anthrax vaccines: past, present, and future. *Vaccine,* 9, 533-539
White, H. D., & McCain, K. W. (1989). Bibliometrics. *Annual Review of Information Science and Technology*, 24, 119-186.